3803ICT Group Assignment

Group Members:

Nathan Cowan- s5143344

Haley Wakamatsu- s5099622

Yasin Cakar- s2921450

Contributions:

Nathan Cowan: Part 1, Part 2, Part 3, Case Study 1 +all tables and graphs therein

Haley Wakamatsu: Case Study 2

Yasin Cakar: No contribution to report at time of submission

Contents

Part 1 - Data Preparation and Preprocessing

Dataset Description

The number of unique variations for the thirteen attributes are:

Id	318477
Title	168065
Company	40629
Date	163
Location	66
Area	20
Classification	31
SubClassification	339
Requirement	234288
FullDescription	250902
LowestSalary	11
HighestSalary	11
JobType	5

Possible values for Jobtype were:

- NULL
- Full Time
- Contract/Temp
- Part Time
- Casual/Vacation

There are no completely null entries.

The columns Id, Title, Date, LowestSalary, and HighestSalary have no null values.

The columns Id, Title, Date, LowestSalary, and HighestSalary have no null values.

Every attribute is typed as a string, except for HighestSalary and LowestSalary, which are typed as a 64-bit integer.

Id gives a unique identifier for each job listing, however these are not in a consistent format, some are integers whilst others are not. This could be normalised in pre-processing.

Title contains the title of the job listing.

Company contains the name of the company that posted it.

Location, and Area, are related categorical attributes. Location represents a larger region, and Area is a more specific place within that.

If present, Classification and SubClassification explain the industry of the job (e.g. a listing for a forklift operator is listed with Classification as “Manufacturing, Transport & Logistics” and SubClassification as “Warehousing, Storage & Distribution”).

The attributes Requirements and FullDescription are not categorical, and appear to have no standard format.

Preprocessing

First the dataset was sampled randomly down to ~10,000 entries, or 1/32 of the original size due to hardware limitations.

Some Id values were at the start of a url-like string, in these cases the trailing characters were dropped and only the numeric Id itself was kept. The remaining string was cast as int64.

The ‘Date’ attribute was retyped into ‘datetime64’.

The Id, Title, Date, LowestSalary, and HighestSalary columns have no null entries. Only these values are considered necessary for an entry, so rows with missing data in other columns were grouped into a new ‘Other’ category (*de facto* a replacement for null*)*, so that they can be excluded from analyses that focus on that column as needed, but the data from the rest of the entry can still be used.

After processing, the number of unique options for Title, Company, Location, Area, Classification, SubClassification, and Jobtype become 8411, 4536, 66, 20, 31, 315, and 5 respectively.

Id	318477	**9928**
Title	168065	**8411**
Company	40629	**4536**
Date	163	**141**
Location	66	66
Area	20	20
Classification	31	31
SubClassification	339	**315**
Requirement	234288	**9681**
FullDescription	250902	**9288**
LowestSalary	11	11
HighestSalary	11	11
JobType	5	5

FullDescription included HTML tags like <p>...</p>, which had to be removed before tokenisation.

Title,Requirement, and FullDescription were each tokenised into new the columns ‘TitleTokens’,’RequirementTokens’, and ‘FullDescriptionTokens’ respectively.

Hypothesis

Given the existence of places famous for a particular industry such as ‘Silicon Valley’, it is expected that the number of jobs, their classifications, and salary will vary greatly by location and others will be more or less constant.

It is also expected that there will be a seasonal aspect to the data, such as contract and part-time jobs being more common near christmas and full-time positions appearing near relevant dates like the financial new year.

Part 2 - Data Analysis and Interpretation

Metadata

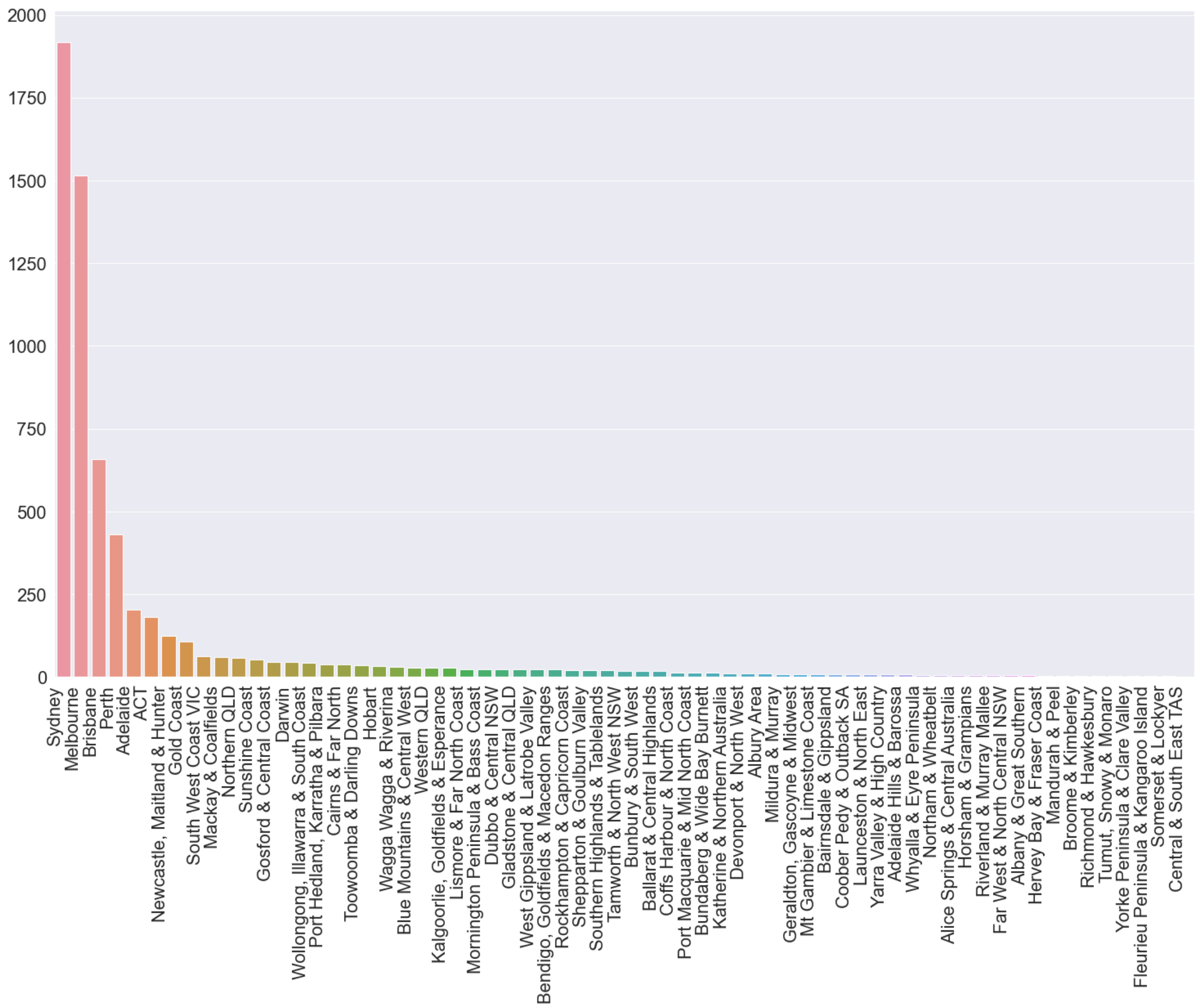
The location of each job is given generally in the 'Location' attribute and more specifically in the 'Area' attribute

The sector of each job is given in the 'Classification' attribute and the sub-sector is given in the 'SubClassification' attribute

For each entry the existing lowest and highest salary attributes are used to calculate a salary range, and average salary.

Market by Location

The market size in each city can be visualised easily by plotting the number of job applications in each location:

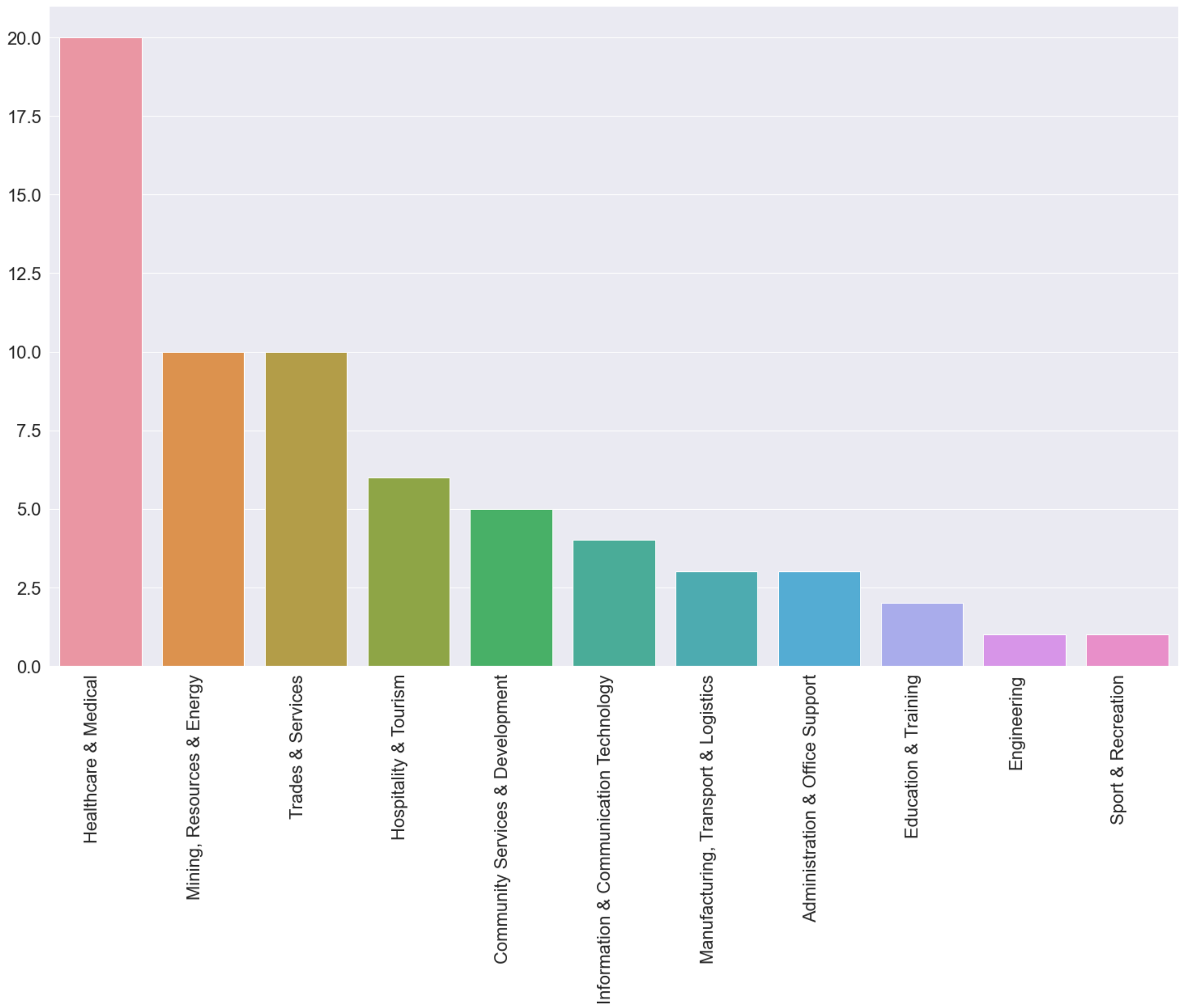


Here it is clear that the capital cities, namely Sydney, and Melbourne account for the vast majority of the job market.

Which Classification is most common in each location can be listed:

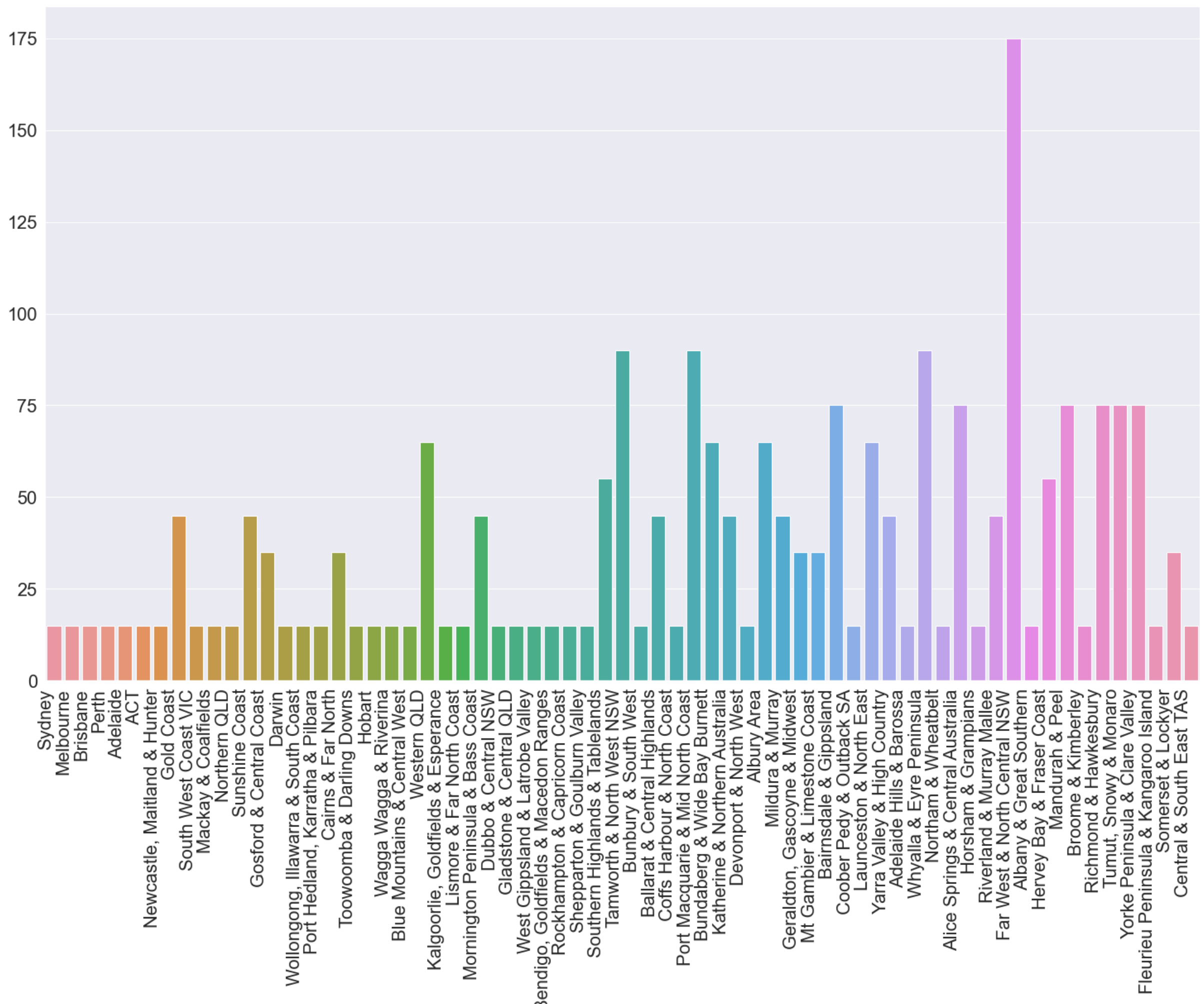
Sydney	Information & Communication Technology
Melbourne	Information & Communication Technology
Brisbane	Information & Communication Technology
Perth	Mining, Resources & Energy
Adelaide	Manufacturing, Transport & Logistics
ACT	Information & Communication Technology
Newcastle, Maitland & Hunter	Healthcare & Medical
Gold Coast	Hospitality & Tourism
South West Coast VIC	Hospitality & Tourism
Mackay & Coalfields	Mining, Resources & Energy
Northern QLD	Trades & Services
Sunshine Coast	Hospitality & Tourism
Gosford & Central Coast	Trades & Services
Darwin	Trades & Services
Wollongong, Illawarra & South Coast	Trades & Services
Port Hedland, Karratha & Pilbara	Mining, Resources & Energy
Cairns & Far North	Healthcare & Medical
Toowoomba & Darling Downs	Healthcare & Medical
Hobart	Healthcare & Medical
Wagga Wagga & Riverina	Healthcare & Medical
Blue Mountains & Central West	Trades & Services
Western QLD	Healthcare & Medical
Kalgoorlie, Goldfields & Esperance	Mining, Resources & Energy
Lismore & Far North Coast	Healthcare & Medical
Mornington Peninsula & Bass Coast	Hospitality & Tourism
Dubbo & Central NSW	Administration & Office Support
Gladstone & Central QLD	Mining, Resources & Energy
West Gippsland & Latrobe Valley	Healthcare & Medical
Bendigo, Goldfields & Macedon Ranges	Healthcare & Medical
Rockhampton & Capricorn Coast	Manufacturing, Transport & Logistics
Shepparton & Goulburn Valley	Engineering
Southern Highlands & Tablelands	Hospitality & Tourism
Tamworth & North West NSW	Healthcare & Medical
Bunbury & South West	Mining, Resources & Energy
Ballarat & Central Highlands	Trades & Services
Coffs Harbour & North Coast	Trades & Services
Port Macquarie & Mid North Coast	Administration & Office Support
Bundaberg & Wide Bay Burnett	Healthcare & Medical
Katherine & Northern Australia	Education & Training
Devonport & North West	Community Services & Development
Albury Area	Healthcare & Medical
Mildura & Murray	Healthcare & Medical
Geraldton, Gascoyne & Midwest	Mining, Resources & Energy
Mt Gambier & Limestone Coast	Community Services & Development
Bairnsdale & Gippsland	Healthcare & Medical
Coober Pedy & Outback SA	Mining, Resources & Energy
Launceston & North East	Healthcare & Medical
Yarra Valley & High Country	Healthcare & Medical
Adelaide Hills & Barossa	Healthcare & Medical
Whyalla & Eyre Peninsula	Trades & Services
Northam & Wheatbelt	Healthcare & Medical
Alice Springs & Central Australia	Community Services & Development
Horsham & Grampians	Healthcare & Medical
Riverland & Murray Mallee	Hospitality & Tourism
Far West & North Central NSW	Mining, Resources & Energy
Albany & Great Southern	Sport & Recreation
Hervey Bay & Fraser Coast	Education & Training
Mandurah & Peel	Mining, Resources & Energy
Broome & Kimberley	Community Services & Development
Richmond & Hawkesbury	Administration & Office Support
Tumut, Snowy & Monaro	Manufacturing, Transport & Logistics
Yorke Peninsula & Clare Valley	Trades & Services
Fleurieu Peninsula & Kangaroo Island	Trades & Services
Somerset & Lockyer	Community Services & Development
Central & South East TAS	Healthcare & Medical

This shows that the top 3 locations all have ‘Information & Communication Technology’ as the most common classification by location, overall the most common job sector by location is actually ‘Healthcare & Medical’, which is visualised more cleanly below:

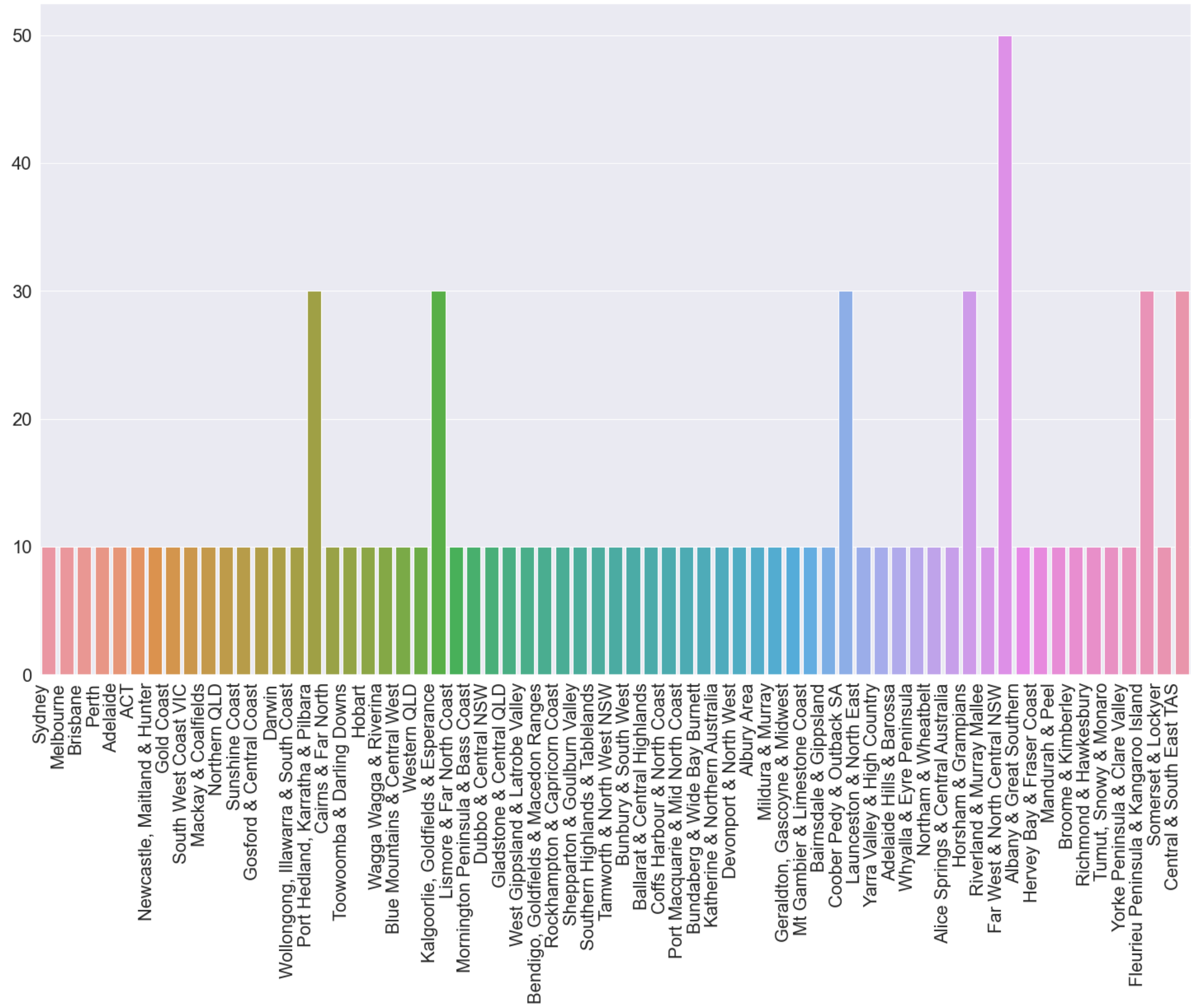


The most common SalaryRange, and AverageSalary for each location can be graphed similarly.

AverageSalary by Location:



SalaryRange by Location:



Together these show that the largest markets are also among those with the smallest SalaryRange and lowest AverageSalary.

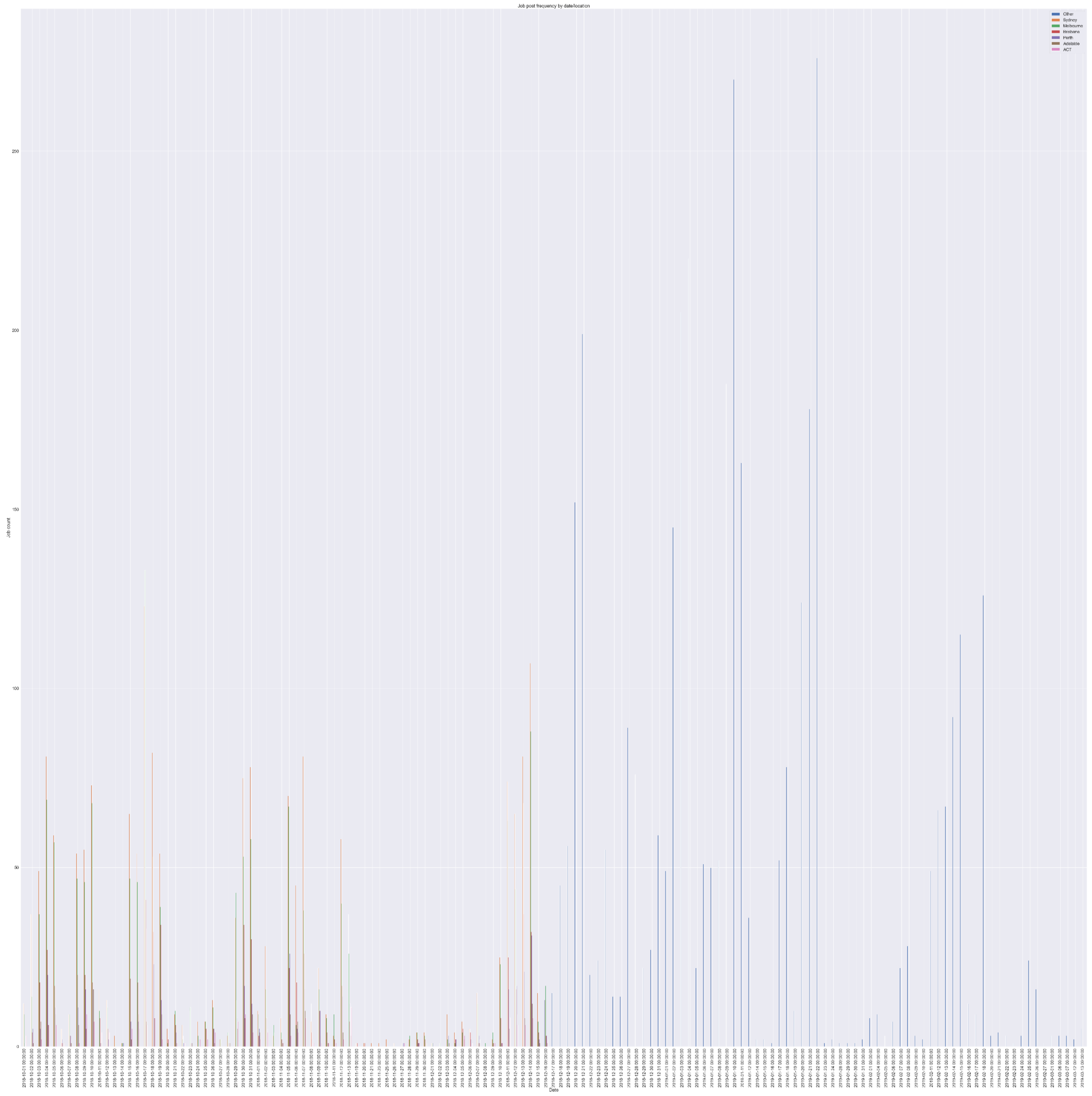
The location with the highest AverageSalary by far is ‘Far West & North Central NSW’, which having so few entries is an extreme outlier and a notable occurrence both.

The most common SubClassification can be listed, but given that both Classification and SubClassification are categorical, little can be learned from comparing them.

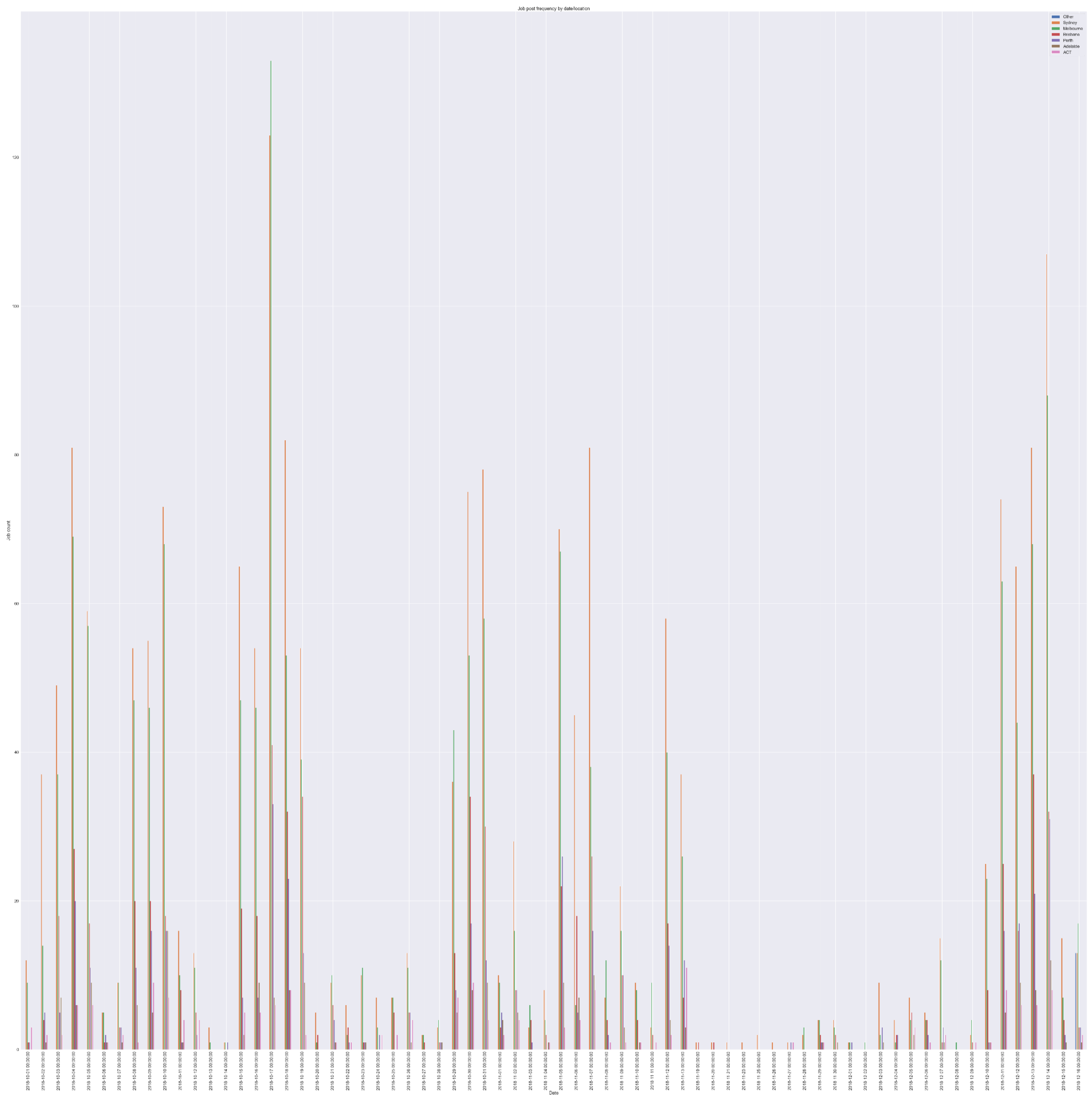
Information & Communication Technology	Developers/Programmers
Trades & Services	Automotive Trades
Healthcare & Medical	Nursing - Aged Care
Hospitality & Tourism	Chefs/Cooks
Manufacturing, Transport & Logistics	Warehousing, Storage & Distribution
Accounting	Financial Accounting & Reporting
Administration & Office Support	Administrative Assistants
Education & Training	Other
Construction	Project Management
Sales	Sales Representatives/Consultants
Retail & Consumer Products	Retail Assistants
Government & Defence	Government - State
Mining, Resources & Energy	Mining - Engineering & Maintenance
Engineering	Civil/Structural Engineering
Community Services & Development	Aged & Disability Support
Legal	Corporate & Commercial Law
Banking & Financial Services	Compliance & Risk
Marketing & Communications	Marketing Communications
Call Centre & Customer Service	Customer Service - Call Centre
Human Resources & Recruitment	Recruitment - Agency
Real Estate & Property	Residential Leasing & Property Management
Design & Architecture	Architecture
Insurance & Superannuation	Claims
CEO & General Management	General/Business Unit Manager
Sport & Recreation	Fitness & Personal Training
Consulting & Strategy	Management & Change Consulting
Advertising, Arts & Media	Programming & Production
Science & Technology	Environmental, Earth & Geosciences
Farming, Animals & Conservation	Horticulture
Self Employment	Self Employment

Useful information for each individual Classification is still available however, such as 'Nursing - Aged Care' being the most common for 'Healthcare & Medical' despite having no real bearing on the others.

Every entry has a date, listing the data by date for each of the top Locations shows:



Looking at this it is shown that the location data is not available after Dec 12. 2018 so the data is sliced again to only show what is known:



There are many significant spikes in job offers.

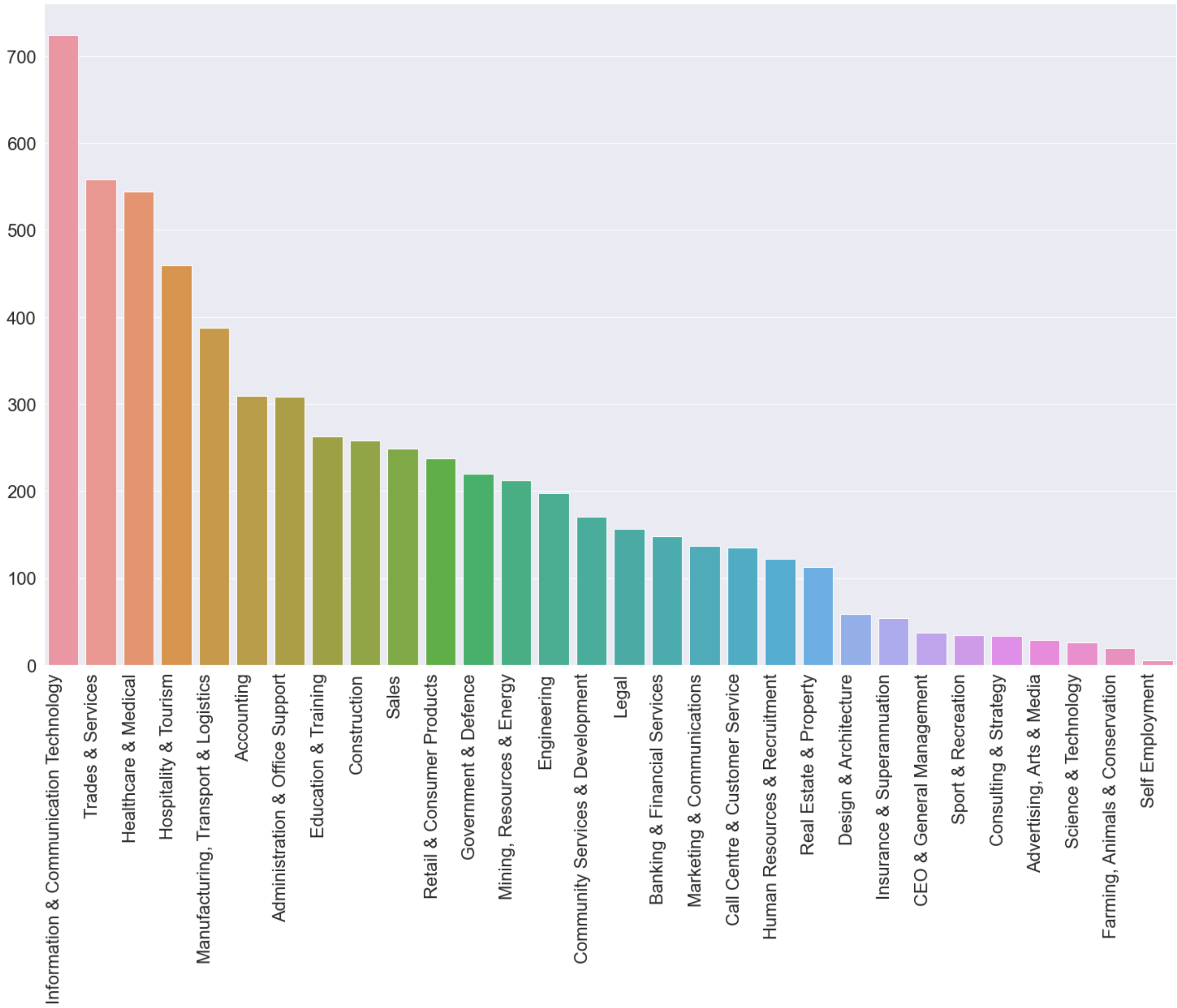
The December spike could be attributed to the beginning of Christmas holidays which usually creates a lot of temporary contract or casual jobs.

Another spike at the end of October is centred around Halloween, although exactly how Halloween affects the job market is not clear.

The largest spike is also in October, centred around the 17th. This can be attributed to a delayed reaction to the new financial quarter starting October 1st, which took businesses two weeks to process and decide what jobs they could offer.

Market by Sector

Just like for Locations, market share can be visualised as frequency by Classification:

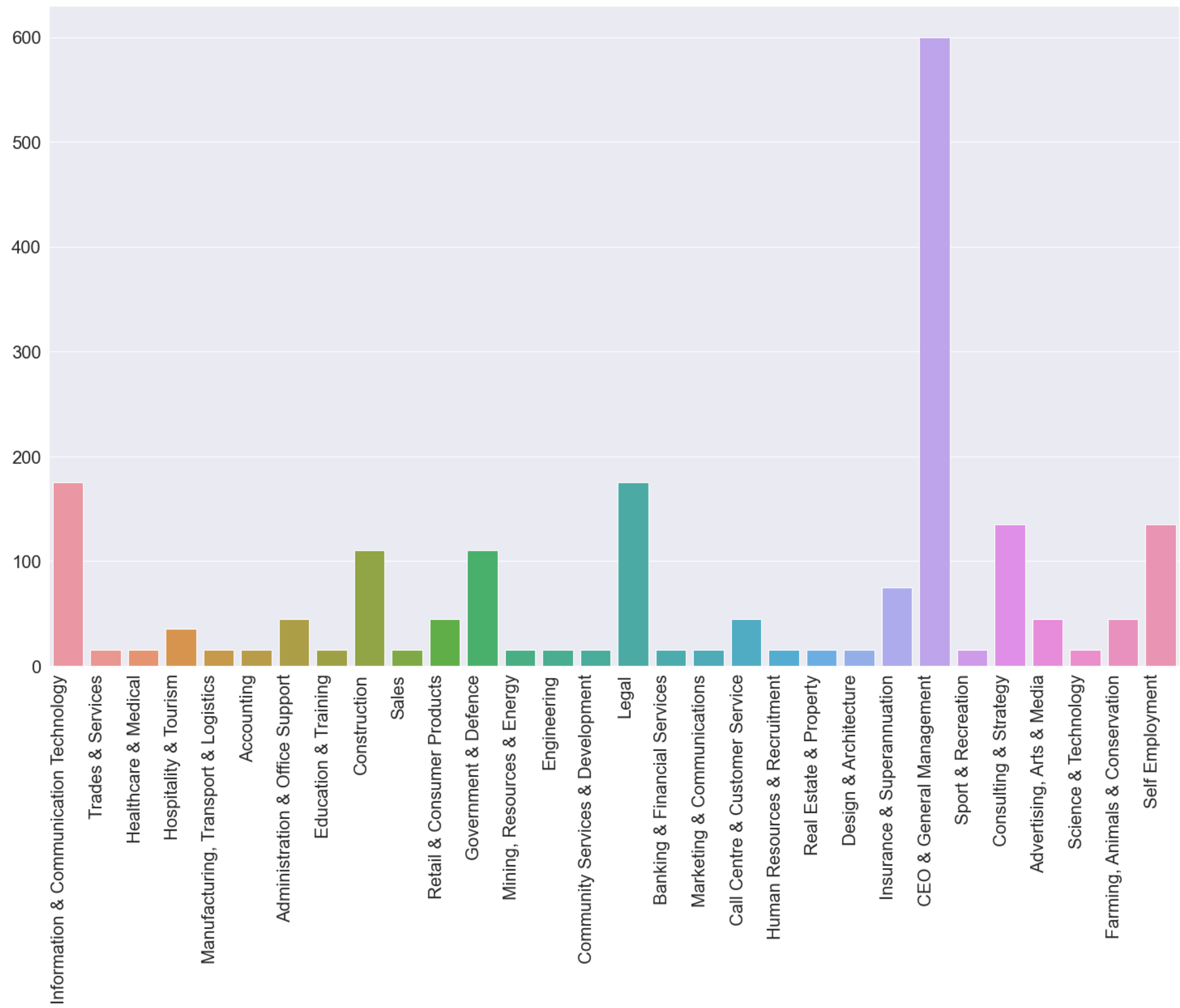


This shows that overall the market is mostly made up of ICT, Trades, Healthcare, and Manufacturing the rest decreasing significantly.

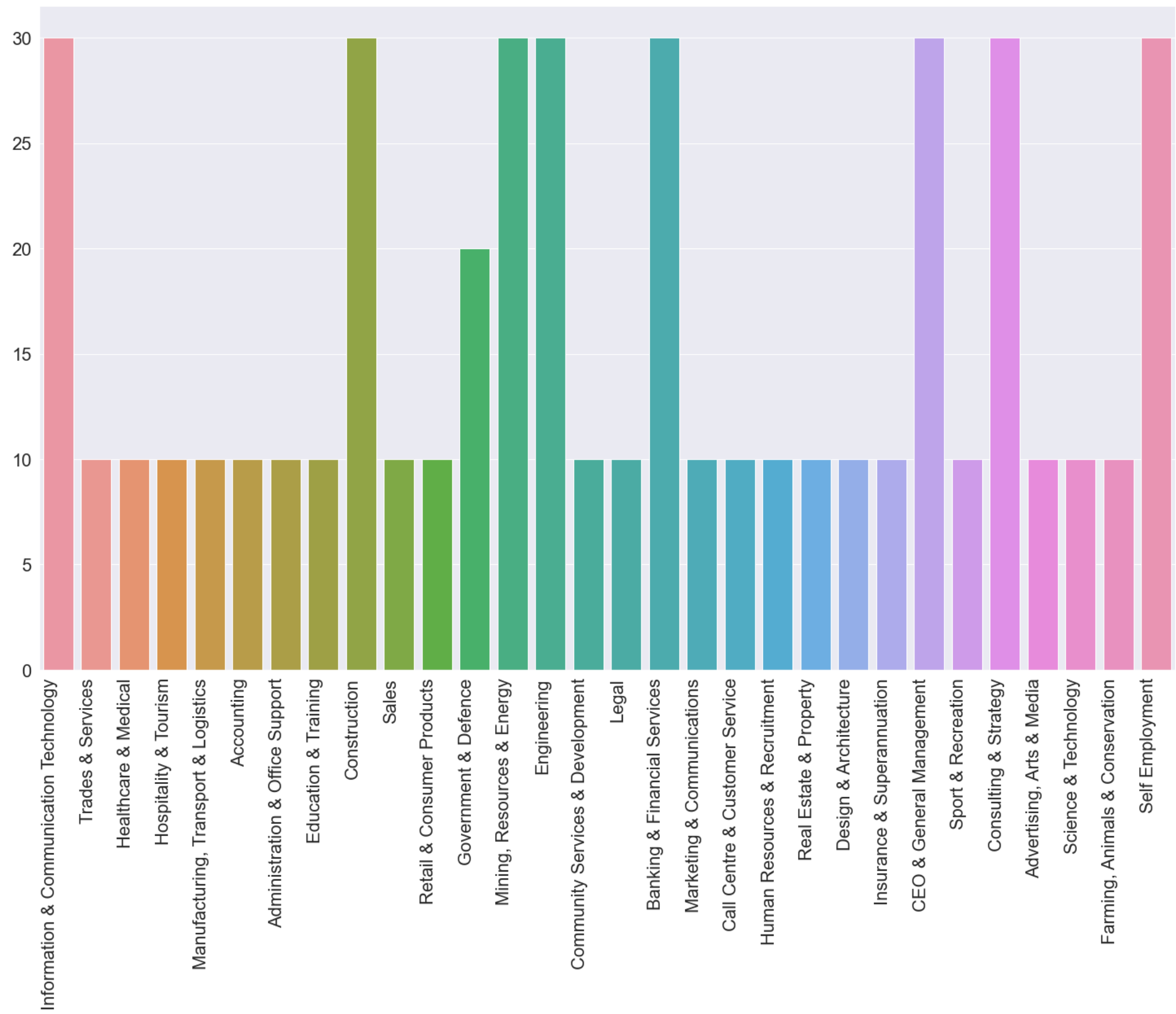
It also shows that job Classifications are much more diverse than job Locations, which is much more focused on the most populous locations.

Along with AverageSalary, and SalaryRange:

AverageSalary:



SalaryRange:



With these plots still ordered according to market share, we can see that there's little correlation between market share and AverageSalary or SalaryRange, and that most Classifications are based on minimum wage with a 10k spread.

This does not have to be true for most actual jobs, but it does appear to be based on frequency analysis.

Part 3 - Evaluation

Findings

Although the most location diverse job sector is ‘Healthcare & Medical’, it is still notably behind ‘Information & Communication Technology’ as the most common job sector.

Additionally, the highest most frequent value for AverageSalary by Classification is neither of those, instead it is ‘CEO and General Management’, far ahead of any others.

Despite this disparity in average pay, the most frequent SalaryRange is actually no higher than 30k for any Classification, including ‘CEO and General Management’.

As far as comparing Classification & SubClassifications it may be possible for a more advanced, and specific analysis to make connections between ‘similar’ Classifications such as the occurrence of ‘Banking & Financial Services - Compliance & Risk’ correlating with ‘Insurance & Superannuation - Claims’, but this is beyond the scope of this report.

Refinement

All of the data is sourced from ‘seek.com’, which could impose a severe bias compared to a dataset with a variety of sources.

The inclusion of ‘Requirement’, and ‘FullDescription’ significantly inflate the amount of processing the dataset requires, for relatively little returns compared to other attributes like, HighestSalary, Location, or Classification as the information these attributes contain is exclusively human-readable.

Websites like ‘seek.com’, and ‘indeed.com’ are useful for collecting the most general information on the job market, however these represent a very limited subset of the job market as a whole, and the data obtained should be compared and/or supplemented by more official information such as from the Australian Bureau of Statistics, which keeps accurate if aggregate information.

Implications

There are clear implications for prospective employees working in particular sectors to know which subsector is the most commonly sought.
More generally for students it is especially relevant which sectors tend to pay better or worse and which are most in demand.

There are also wider implications for both employees and employers in identifying which sectors have the highest AverageSalary, and SalaryRange. According to the analysis of this dataset higher pay is highly correlated to management positions of any kind- from C-level positions to a project manager.

Part 4

Case Study 1

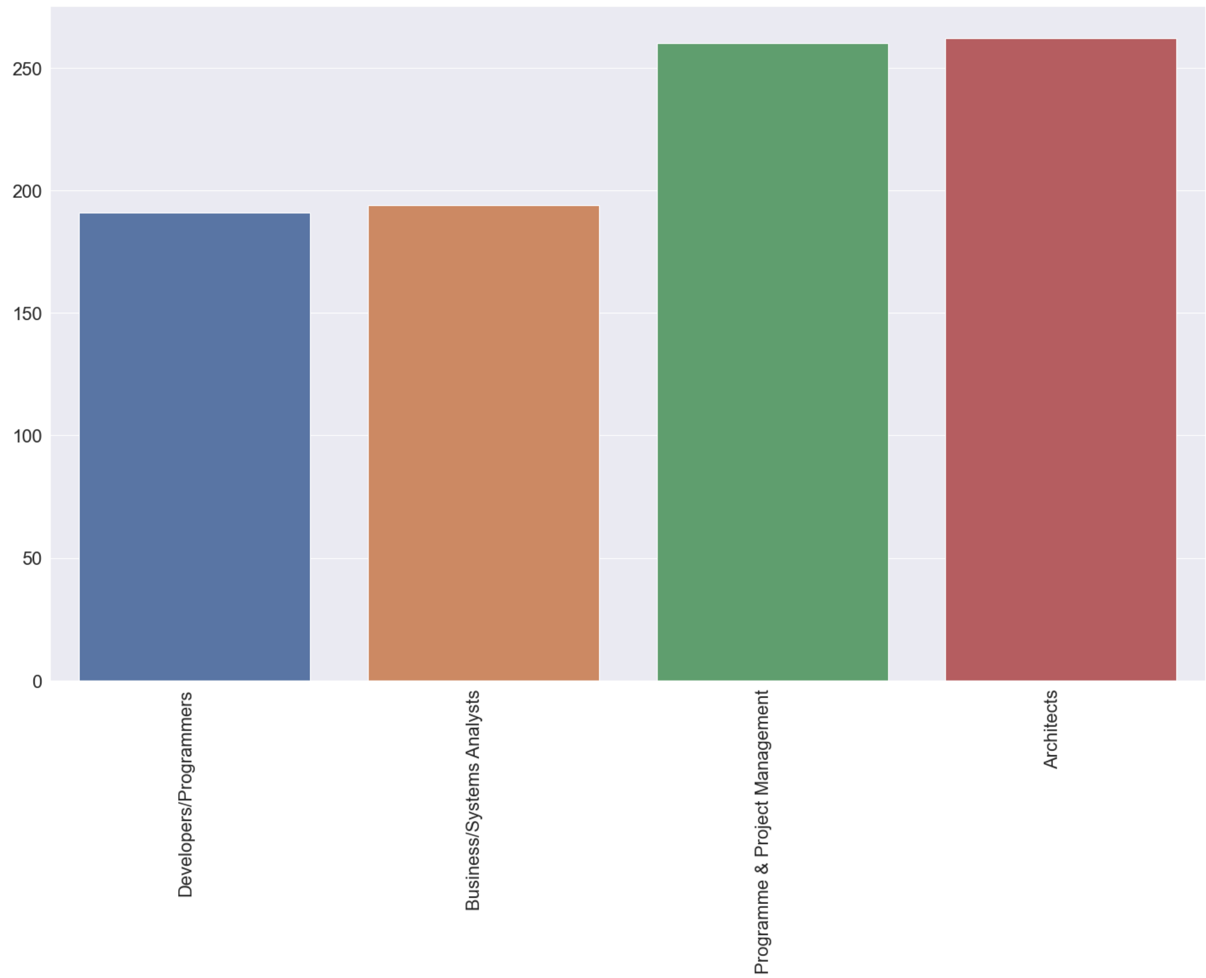
The first case study is to advise Matthew, a computer science student in his first year of study, on which courses and skills would be most beneficial to him becoming an IT expert, based on current market data.

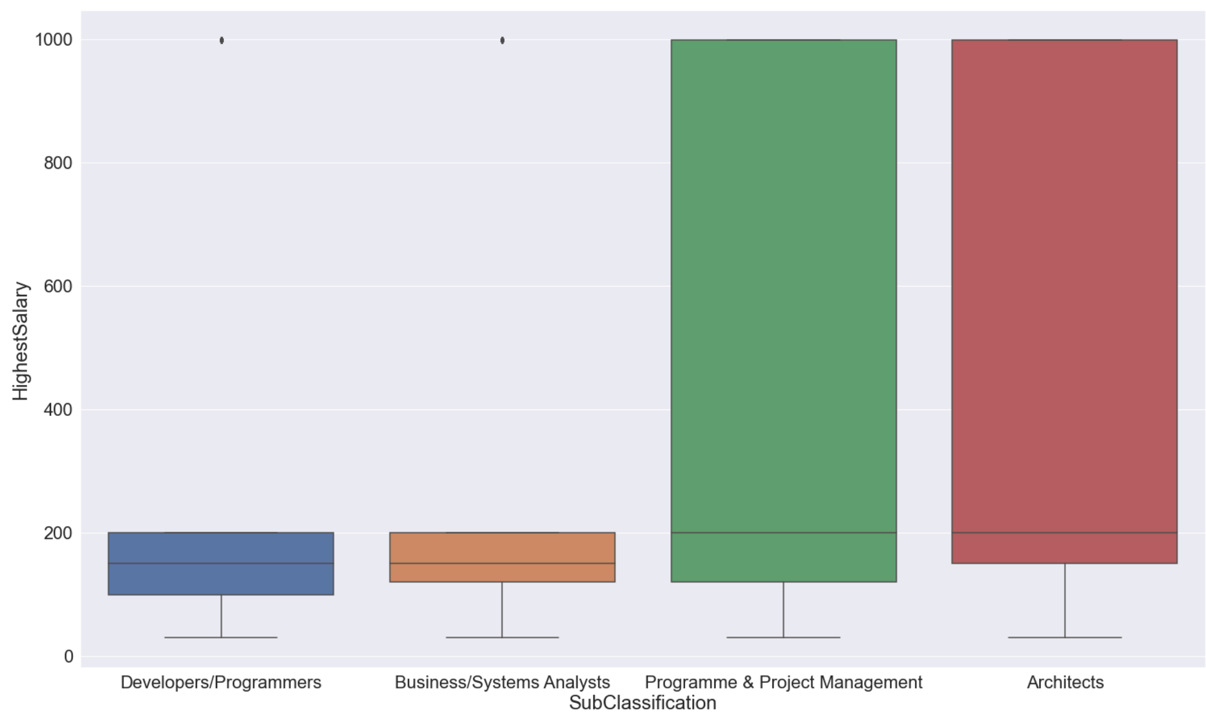
Given that Matthew is a computer science student, the dataset can be easily limited to only the 724 job listings where Classification equals ‘Information & Communication Technology.

There are 20 unique values for SubClassification. The most frequent of these, being ~160% as frequent as the next, is ‘Developers/Programmers’. It isn’t surprising that a computer science student should learn programming. The less common SubClassification values are ‘Business/Systems Analysts’, ‘Programme & Project Management’, and ‘Architects’ (one who designs computer systems, not buildings) which account for an aggregate ~50% of the data.

Also to be considered beyond simple employability is payment. Are there any significant differences in payment between these four options?

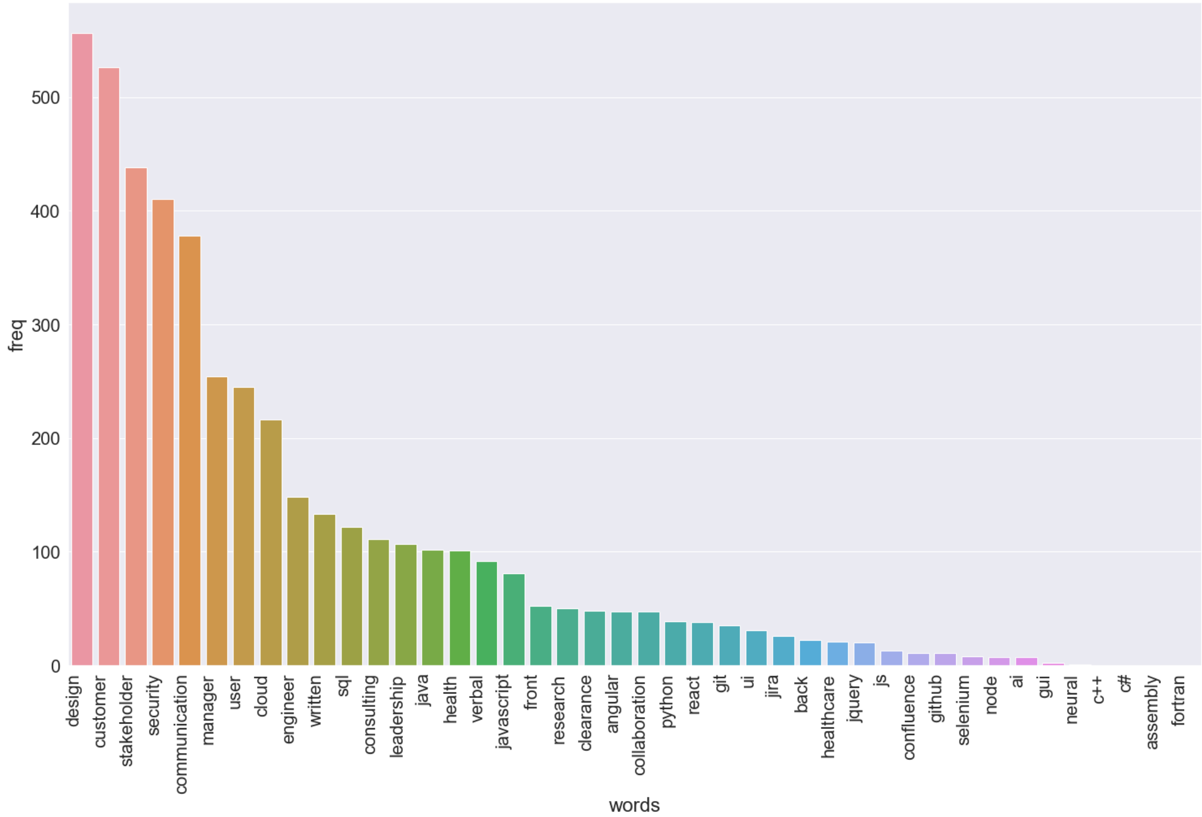
The average of the lowest and highest salary, as well as the highest salary, within each category is compared below:





Looking at the graphs, we can see that developers and analysts are paid very similarly, but there is an enormous difference in the pay ceiling between them and managers and architects. Despite this, the average pay has a much smaller difference of about $70,000.

Project managers jobs are a more common offering than architect jobs, but what actual skills does an IT expert actually require? The relative frequency of words between the two groups is shown below.



The relative frequency of keywords shows that ‘customer’, ‘stakeholder’, ‘security’, ‘communication’, and ‘user’ are all frequently mentioned. Specifically, communicating with customers or stakeholders seems to be a skill in high demand.

In conclusion, not all the skills most desired by employers are universal to computer science graduates. I would recommend that Mathew choose courses and electives that focus on identifying user needs and IT-customer communication.

Case Study 2

Since we need to deal with open-ended, qualitative data instead of clear-cut discrete categories, it would be easiest for everyone to attempt natural language processing, possibly on someone’s CV, LinkedIn profile, or other sources such as recommendation letters.

This is a great place to use a recommendation system based on word frequency and textual similarity, such as TF-IDF. The data also provides location data of the workplace, which can let the clients only browse for jobs within a certain radius of a specified location (or maybe, all train stations, or along a specified walkable path).

However, assuming that some manual data input and/or verification can be used, each client’s résumé and job listing can be manually tagged with the necessary skills, such as programming languages or tools (e.g. C++, NodeJS/AngularJS, Heroku) and job content (e.g. customer liaison/communication, project organisation, UI design). That way, the data can be organised in a more accurate, human-readable way.

Both solutions can be used in tandem, where search results based on textual similarity can be filtered by these manual tags. At the end of the day, no one solution is the only correct solution, and the best decision is usually the most balanced one.